Empir Econ (2017) 53:1681–1710
DOI 10.1007/s00181-016-1169-2

Improving the prediction of ranking data

Marco A. Palma[1]

Received: 22 April 2015 / Accepted: 4 August 2016 / Published online: 22 September 2016

Abstract By using the same number of alternatives for every respondent, all ranking elicitation methods in the literature including full, partial, and best–worst rankings assume respondents know and are able to rank *the same* number of alternatives. A simple survey elicitation mechanism allowing for individual heterogeneity in the number of rankings for ranked-ordered data is proposed. Using the proposed ranking mechanism as a data augmentation tool yields higher prediction of ranking choices compared to conventional rankings and best–worst methods. The results provide robust evidence of differences in error variance scale and the structure of the underlying utility preferences across ranking stages, including best–worst rankings. The highest predictive power was achieved with the proposed ranking method using only the best ranked alternative. Including any additional rankings other than the best alternative reduces predictive power. Nevertheless, if more than one ranking is used to model preferences, then better predictions are achieved by using the top two best ranked alternatives as supposed to the exploded best–worst rankings. The results stand as a warning about equating ranking choices to true underlying utility preferences across different ranking elicitation stages or mechanisms without properly testing for symmetry and stability of preferences.

Keywords Best–worst · Error variance · Random parameters · Scale parameter · Stability of preferences

JEL Classification C83

✉ Marco A. Palma
mapalma@tamu.edu

[1] Department of Agricultural Economics, Texas A&M University, 2124 TAMU, College Station, TX 77843, USA

1 Introduction

Consumers make many decisions every day. They decide which products to buy and where to buy them. In a real purchasing setting, consumers are confronted with dozens and sometimes even hundreds of choices among similar competing products. In order to replicate the consumer decision process, researchers often use primary data to analyze consumer choices relying on surveys to elicit preferences. Many surveys elicit individual preferences by asking respondents to rank several competing alternatives. Given a set of alternatives, respondents may be asked to provide a rank-order of their preferred brands of sodas, or their preferred football teams or their preferred transportation mode to work. The traditional full ranking procedure involves the selection of the most preferred option in a choice set with J alternatives $C = \{1, 2, 3, 4, \ldots, J\}$, eliminating it from the choice set, and then selecting the most preferred option out of the remaining alternatives and rank it second; this process is repeated again with the remaining choices until all the alternatives are ranked. Statistical procedures are then used to analyze the ranking choices based on the random utility theory (McFadden 1974).

In order to estimate preferences on the most preferred alternative, a traditional multinomial logit (MNL) or conditional logit (CL) model can be used. The MNL produces estimates for individual specific characteristics for each alternative minus one which can lead to over-parameterization in models with large number of choice probabilities. The CL refers to choice-specific attributes. Other than that both models are mathematically equivalent (Greene 2012). Both the MNL and CL models provide preferences for the most preferred option only, and no additional information is obtained even if other preference rankings are available. However, in many cases, researchers use elicitation procedures to obtain full or partial rankings. Beggs et al. (1981) introduced the rank-order logit (ROL) model for the analysis of preference rankings with an application to the potential demand for electric cars. The ROL model allows for the analysis of individual preferences for all (or partial) rankings when ranked-ordered data are available. This model has become the ranking model of choice in the literature and has been applied to many fields, including socioeconomics and politics (Koop and Poirier 1994), music (Ophem et al. 1999) and marketing (Bronnenberg and Vanhonacker 1996) to name a few.

The ROL is useful as a data augmentation tool for small sample sizes by allowing the data to include more observations. However, there are several reasons why one could argue that using the full set of rankings in the choice set C may not be appropriate. For example, if the available alternatives for transportation mode to work are car, bus, bicycle and train, one or more alternatives, such as train, may not be available to all respondents in some cities or in his/her particular route to work. Relevant alternatives may not always be so obvious; in the case of the preferred brands of sodas, some respondents may not know all the brands in the choice set, making the ranking of the unknown or unfamiliar alternatives problematic. Furthermore, when ranking preferred football teams some alternatives may be viewed positively, while others may be viewed with indifference or even negatively. *Relevant alternatives* for the purpose of this article are defined as those alternatives in the consideration set that an individual is capable of ranking. Note that in this context, the number of relevant alternatives

in the consideration set may differ for each respondent. That is, for one respondent maybe only the top-ranked alternative is relevant, while other respondents may have two or more relevant alternatives in the consideration set. Hausman and Ruud (1987) developed specification tests to compare the consistency of top (best) ranked choices with the lower (worst) ranked choices. They found that respondents tend to rank more carefully the top-ranked choices and suggested using a subset of the choice set which includes only a few of the top rankings. Because their results depended on the particular products being studied and the total number of alternatives being ranked, they concluded there were no general rules a priori on survey design to address the appropriate number of ranks to be used in the analysis. Fok et al. (2012) argue that respondents may not perform the ranking task according to their true preference. They point out that even if respondents know their true preferences, they may find the task too tedious or complicated.

Chapman and Staelin (1982) found that as the number of rankings used in the analysis increases, and hence the number of independent choice observations increases, there is a reduction in the sampling variance. They also noted that using more rankings means using more observations with 'lower ranks,' which according to them are more likely to be characterized as random or noisy. They concluded a trade-off exists between the number of rankings used and the associated reduction in variance with the reliability of the parameter estimates. In order to solve this problem Chapman and Staelin (1982) suggested using only the first few top-ranked alternatives following a set of rules based on a pooling test for equality of the parameter estimates for different ranks. In doing so, they test the stability of the parameters across ranking stages. Hausman and Ruud (1987) analyzed a set of eight choices to rank preferences for cell phones using full and partial ROL models. They found that as the number of rankings is increased, the absolute value of the estimated coefficients falls. These results are in agreement with the findings of Chapman and Staelin (1982). Hausman and Ruud (1987) proposed a heteroscedastic ROL model that allows the variance in the lower rankings to differ from the variance in the higher rankings by introducing a scale parameter. Recently the scale parameter approach has been used in several applications to account for potential asymmetry of preferences across different ranking stages (Scarpa et al. 2011; Auger et al. 2007; Flynn et al. 2007; Cohen et al. 2009).

Obtaining and using more ranking information is becoming more important, particularly given the rise of individual heterogeneity applications in the literature (Hess and Rose 2009). The elicitation of full (or partial) rankings is designed with the purpose of improving the precision of the parameters and the prediction of the models (Vermeulen et al. 2011). If a partial number of rankings are used, the problem is identifying the optimal number of rankings to use given a set of J available alternatives. Several studies have estimated partial ranked-ordered logit (PROL) models. In general, the procedure involves a multi-step cognitive choice process in which respondents first identify a subset of products within all the alternatives and then evaluate those choices. In their study of brand preferences, Bronnenberg and Vanhonacker (1996) suggest a two-stage process involving a choice set formation of relevant brands, and then a selection process from the alternatives in the choice formation stage. Chiang et al. (1998) proposed an integrated consideration set-choice model that accounts for

heterogeneity in the first stage of the consideration set and the second step of the choice formation model using a nonparametric Markov chain Monte Carlo (MCMC) sampling procedure applied to scanner panel data. Ophem et al. (1999) suggest a three-step process for selecting the number of relevant rankings. The first step is to choose the $0.5J$ most preferred choices from J alternatives. The second step is to select $0.25J$ most preferred alternatives from step one. The third step is to rank the alternatives in the choice set from step two. Hensher and Ho (2014) evaluated the behavioral effects of choice set formation of relevant alternatives and whether adding certain alternatives improves the prediction of the most preferred alternative. They find differences in preferences depending on which alternatives are included in the relevant choice set. In terms of evaluating the attractiveness or acceptability of different alternatives in the consideration set, Hensher and Rose (2012) found improvements in prediction when accounting for the acceptability of each alternative.

An alternative method for partial rankings consists of eliciting the best and worst (BW) alternatives in a choice set (Finn and Louviere 1992). The BW method rose as a way to facilitate the cognitive process of ranking competing alternatives. The general idea is that asking respondents to select extreme options and moving away from middle ranked alternatives results in more consistent and stable preferences (Flynn et al. 2007; Giergiczny et al. 2013). There are three relevant BW cases (Louviere et al. 2013; Rose 2014). The object case (case 1) consists of participants being asked to select the best and worst objects, and no information is provided about the attributes of each object (Finn and Louviere 1992). In the profile case (case 2), participants are asked to evaluate combinations of attributes, and select the best and worst attribute levels for each alternative (Rose 2014). In the choice alternative case (case 3), participants choose the best and worst designed profiles from several choice sets derived from an experimental design (Marley and Pihlens 2012; Giergiczny et al. 2013).

All of the partial ranking methods described above rely on adding *the same* number of choices per respondent, thus assuming that all respondents know their utility and are able to rank *the same* number of alternatives. In reality, this assumption may be too restrictive, as some participants may only care about the best available alternative, while others may be interested in more than one alternative. The objectives of this article are: (1) to introduce a simple survey elicitation method for relaxing the homogeneity in the number of rankings assumption and allowing each respondent to have a different number of rankings; (2) to empirically compare the results of full and partial ranked-ordered logit models with and without individual heterogeneity in the number of rankings; (3) to estimate (case 1) BW models of the exploded data; (4) to evaluate the predictive power of the partial ranking methods, including exploded BW, with constant and heteroscedastic error variances; and (5) to test the stability of preferences across ranking stages. The paper is organized as follows. First, the general rank-ordered logit model is presented. Second, a simple survey elicitation mechanism to allow for heterogeneity in the number of relevant rankings is introduced. Third, estimation methods for an empirical application are presented. The results are followed by a summary and conclusions section.

2 The rank-ordered logit model

Following (Chapman and Staelin 1982), a consumer n ($n = 1, 2, 3, \ldots, N$) has a choice set C consisting of J alternatives ($1 < J < \infty$). The alternatives in the choice set are described in terms of their attributes X. Each participant chooses an alternative i in the choice set to maximize utility. The utility function U, which measures the unobserved desirability or attractiveness of an alternative, can be described as $U_{nit} = V_{nit} + \varepsilon_{nit}$, where U_{nit} is assumed to have a deterministic component V_{nit} and an error term ε_{nit}, which is assumed to follow some distribution function. Let the deterministic portion of the utility function for participant n, alternative i in ranking situation t be explained by:

$$V_{nit} = \beta_{nt} x_{nit} \tag{1}$$

where X_{nit} is a vector of product attributes describing alternative i faced by respondent n in ranking situation t, where t is a baseline ranking round and a tasting treatment round.

An individual n would choose alternative i at time t from the choice set $C \Leftrightarrow U_{nit} > U_{njt}$ for $\forall j \neq i$. Let $y_{nit} = 1$ represent that for respondent n the most preferred alternative is i, which implies that $U_{nit} > \max\{U_{n1t}, \ldots U_{njt}\} \forall j \neq i$. Then for an observed ordinal ranking of the choices the probability of being selected is given by:

$$\Pr\left(U_{n1t} > U_{n2t} > \cdots > U_{njt}\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{U_{n1t}} \int_{-\infty}^{U_{n2t}} \cdots \int_{-\infty}^{U_{njt}} \mathrm{d}G\left(U_{n1t}, U_{n2t}, \ldots U_{njt}\right) \tag{2}$$

where $\mathrm{d}G(U_{n1t}, U_{n2t}, \ldots U_{njt})$ is the joint distribution of the U_{nit}'s induced by the error term. Note that throughout this article the most preferred option is assigned a rank of 1, the second preferred option a rank of 2, and so on. Therefore, a preferred (higher) ranked alternative implies a lower numerical value. Theoretically, any joint distribution function can be assumed in Eq. (2). If the error term is assumed to be identically and independently distributed (iid) extreme value over respondents and alternatives, then the probability of alternative i being selected is:

$$\Pr_{nit} = \frac{\mathrm{e}^{(V_{nit})}}{\sum_{j=1}^{J} \mathrm{e}^{(V_{njt})}} \tag{3}$$

where J refers to the total number of alternatives in the choice set for respondent n at time period t. Data on the most preferred option are enough to estimate the model; however, an efficiency gain can be obtained by asking respondents to rank more alternatives in the choice set. Then the probability of observing the rank ordering for person n and alternative i at period t is:

$$\Pr\left[U_{nit1} > U_{nit2} > \cdots > U_{njt}\right] = \Pr_{nit} = \prod_{j=1}^{J-1} \frac{\mathrm{e}^{(V_{nit})}}{\sum_{j=1}^{J} \mathrm{e}^{(V_{njt})}} \tag{4}$$

The rank-order data is 'exploded' to become a series of independent choice decisions where the most preferred alternative (best) is selected first and removed from the choice set; next, the second most preferred alternative (second-best) out of the remaining alternatives is selected, and so on, until all products are ranked. In reality, respondents do not have to make the decision process in that order, but this provides a practical way of visualizing the concept of Eq. (4).[1] Suppose that participant n faces the choice set $C = \{a, b, c, d\}$ and selects the full ranking ordering $b \succ a \succ c \succ d$. This would imply that for that individual, $U_n(b) > U_n(a) > U_n(c) > U_n(d)$. The data is exploded to three choice decisions where the selected alternatives are removed from the choice set at each step and the probability of the exploded rank-order is:

$$\Pr(b > a > c > d) = \frac{\mathrm{e}^{(V_{ntb})}}{\sum_{j=a,b,c,d} \mathrm{e}^{(V_{ntj})}} \cdot \frac{\mathrm{e}^{(V_{nta})}}{\sum_{j=a,c,d} \mathrm{e}^{(V_{ntj})}} \cdot \frac{\mathrm{e}^{(V_{ntc})}}{\sum_{j=c,d} \mathrm{e}^{(V_{ntj})}} \quad (5)$$

The conditional distribution of the ordering is independent of the ranking because of the independence of irrelevant alternatives property (IIA) of the multinomial logit model, which is also assumed by the ROL model (Beggs et al. 1981).

2.1 A ROL model with partial rankings

The full ranked-ordered logit model assumes that respondents know their utility for each alternative and are able to rank each alternative according to the framework. However, as Fok et al. (2012) point out this assumption may not always hold. Respondents may not be familiar with all the available alternatives in the choice set or may consider some alternatives less important and commit less time or effort. Even though including more alternatives would reduce the variance, it would also bias the parameter estimates toward zero (Chapman and Staelin 1982). Current application methods for partial ranked-ordered logit models consist of estimating separate models for the top j alternatives and merging the data sets based on pooling tests. In the example above, if only the best option is selected, then the probability would be defined as:

$$\Pr(b > a > c > d) = \frac{e^{(V_{ntb})}}{\sum_{j=a,b,c,d} e^{(V_{ntj})}} \quad (6)$$

If the top two options are selected then the probability becomes:

$$\Pr(b > a > c > d) = \frac{\mathrm{e}^{(V_{ntb})}}{\sum_{j=a,b,c,d} \mathrm{e}^{(V_{ntj})}} \cdot \frac{\mathrm{e}^{(V_{nta})}}{\sum_{j=a,c,d} \mathrm{e}^{(V_{ntj})}} \quad (7)$$

[1] The exploded logit model is not reversible. Assuming alternatives are ranked from best to worst would produce different parameter estimates than assuming the ranking was performed from worst to best which is inconsistent with random utility theory.

In this example, selecting the top three alternatives will result in a full ranking ordering as described in Eq. (5) since choosing an alternative in the third step identifies the last ranked alternative.

For the exploded BW ranking procedure, respondents make a sequence of choices by selecting the best alternative first and then out of the remaining choices they select the worst alternative. Depending on the experimental design, this process may be repeated several times. The utility function associated with the worst alternative is then multiplied by -1, assuming that worst is the direct opposite of best (Giergiczny et al. 2013; Rose 2014). That is the characteristics that make the best alternative attractive, make the worst alternative unattractive. The BW partial ranking procedure in the example above would change the outcome probability to:

$$\Pr(b > a > c > d) = \frac{e^{(V_{ntb})}}{\sum_{j=a,b,c,d} e^{(V_{ntj})}} \cdot \frac{e^{(-V_{ntd})}}{\sum_{j=a,c,d} e^{(V_{ntj})}} \tag{8}$$

Note that all partial ranking methods discussed so far assume the same number of choices for each participant. If the best two alternatives are used, then they are used for every respondent. If the exploded BW procedure is used, the ranking stages are the same for every respondent.

2.2 The proposed ranking mechanism

The proposed ranking elicitation mechanism consists of two phases. In phase one, participants choose the alternatives they consider relevant to be included in their consideration set. During the second phase, each participant ranks the alternatives in the consideration set. Hence, each participant would have a different consideration set that may include none, a subset, or all of the alternatives. In order to test and compare the proposed model with traditional full and partial ranking models, an additional alternative was included in the choice set $C = (J + 1)$ to be used as a natural self-selected break point to establish a threshold of *relevant* alternatives in the choice consideration set. The $(J + 1)$ alternative is an 'opt-out' option for 'no-product,' which implies a respondent preferred none of the available alternatives over the option of choosing any of the remaining alternatives in the choice set. By ranking the alternatives respondents would purchase, the elicitation mechanism provides a closer representation of the process a consumer faces in a real purchasing setting. In the context of the example above, in addition to the four alternatives, a, b, c, and d, each participant would evaluate the choice set $C = \{a, b, c, d,$ 'no-product'$\}$. A ranking of $\{b, a,$ no-product, $c, d\}$ for individual n would imply that $b \succ a \succ (np) \succ c \succ d$. The probability of the rank-order for that particular ordering will be that represented on Eq. (7). The 'opt-out' option is not directly used in the analysis, but it is only used to classify the relevant alternatives (i.e., those ranked higher than no-product) and restrict the consideration set to only those relevant alternatives. The opt-out option acts as a natural self-selected endogenous breakpoint in the choice set allowing each respondent to choose the number of alternatives in their consideration set, thus relaxing the assumption that every respondent has *the same* number of alternatives in the consideration set. Behaviorally,

this modeling framework implies that the consideration and ability of subjects to rank a set of alternatives is not the same.[2] The survey implementation procedure proposed would generally only ask participants to rank the alternatives in the relevant consideration set (i.e., alternatives b and a in the example above). Rankings of alternatives below the opt-out option were only elicited in the survey for testing whether the utility preference parameters are symmetric and stable across the ranking stages. If the utility parameters are not stable across the ranking stages, then it can be concluded that a mismatch between the ranking response format and the true preference exists. This framework can be tested by evaluating error variance scale differences and the structural stability of the parameters across the two ranking stages (alternatives ranked below and above the no-product option). The generalized probability of any observed outcome is the expression in Eq. (4) $\forall j \neq i$, s.t. $\text{rank}(j) \succ \text{rank}(\text{no product})$.

2.3 Testing symmetry and stability of parameters across ranking stages

Differences in error variances can be tested by estimating heteroscedastic error variance models through a scale parameter across ranking stages. A statistically significant scale parameter implies asymmetry in variance across the ranking stages. While differences in the variance across ranking stages can be easily accommodated by the scale parameter, a more important test is whether the different ranking stages represent the same utility preferences. If structural differences of the parameters across ranking stages are found, it would provide evidence of non-stable preferences across the two stages (i.e., preferences from choosing alternatives ranked below the opt-out option are different than preferences from alternatives ranked above the opt-out option). The stability of particular ranking stages can also be tested; that is a test of stability of preferences across the best and second-best rankings or best and worst rankings or any other partial rankings. If structural changes in the parameters are found across rankings then the common practice of pooling ranking data sets without properly testing would be inappropriate. This would imply that the set of decision rules to define preferences are not the same across ranking stages. Structural breakpoints are tested by estimating separate models across ranking stages and conducting a test of the pooling hypothesis (Greene 2012). An alternative test conducted for robustness consists of adding to the pooled model the original parameters multiplied by a stage specific indicator variable (Allison and Christakis 1994; Giergiczny et al. 2013).

2.4 Predicting ranking choices

Comparing the results across the estimated models with different number of rankings presents a problem because the models are non-nested and have different number of observations. Since the pseudo-R^2 is a likelihood-based test, comparison across non-nested models is meaningless. The models are compared based on their predictive power. The basic notion used is that a good model is one that predicts choices well. In

[2] A referee pointed out that having different number of alternatives ranked above and below no-product is endogenous for different rankings and may induce asymmetry.

that context, Tjur (2009) recently proposed a measure of fitness for logistic regression models called the coefficient of discrimination. The coefficient of discrimination is defined as: $D = E[\hat{\pi}_1] - E[\hat{\pi}_0]$. It measures the difference of the means of the predicted values for successes and failures, and hence the name, since it measures the model's ability to discriminate between successes and failures. Tjur's statistic was developed for binomial responses, and it compares the mean of the predictions for $y = 1$ (success) with the mean of the predictions for $y = 0$ (failures). The general idea is that a good model (one that predicts well) will have high predicted values for observations when the alternative was chosen and low predicted values when the alternatives were not chosen. With multiple alternatives in the exploded data, the Tjur R^2 was calculated as the difference between the mean of the predicted probabilities when $y = 1$, and the mean of all other alternatives not chosen ($y = 0$). The test is asymptotically equivalent to several R^2 measures used for linear models (Tjur 2009). Since Tjur's R^2 is not based on the likelihood function of the models, it can be used to compare the explanatory power for non-nested models with different number of observations. However, Tjur's R^2 was developed for binary logit regressions and has not been applied no multinomial analysis. For robustness, a compliance test for the predicted probabilities is included (Hensher and Ho 2014). The compliance statistic is the percentage of cases where the chosen alternative had the highest predicted probability. While the compliance test indicates the accuracy of prediction of the chosen alternative, the Tjur's R^2 provides an indication of the distance of the mean predicted values for the chosen and not chosen alternatives.

2.5 The random parameters rank-ordered logit

The availability of multiple observations from the same individual provides rich data to model respondents preference heterogeneity using a random parameter rank-ordered logit (RPROL) model (Train 2009; Hess and Rose 2009). In the RPROL, the stochastic terms are used to represent deviations from the mean coefficients; then the errors can be allowed to be correlated across ranking alternatives. Within the same framework, $U_{nit} = V_{nit} + \varepsilon_{nit}$, for participant n and product alternative i, at time period t, each ε_{nit} is assumed to be iid extreme value over time, respondents, and alternatives. Let j correspond to each of the product options to be ranked, with $J-1$ choice decisions to be made and t treatment rounds, the model can now be specified as: $U_{nit} = \beta_n x_{nit} + \varepsilon_{nit}$, where $V_{nit} = \beta_{nt} x_{nit}$. In order to allow correlation across multiple ranking individuals, the utility function can be modified with further specification of β_n. The β_n is the unobserved vector of coefficients for each individual n that is randomly distributed with a conditional probability density function given by $f(\beta_n|\theta^*)$ where θ^* represents the true parameters of the distribution (Calfee et al. 2001); The stochastic source of error ε_{nit} remains uncorrelated with β_n, and x_{nit} is iid extreme value. The β coefficient vector takes on the form: $\beta = b + \eta_n$ where b is the population mean and η_n are individual deviations from the average population (Calfee et al. 2001). Utility is specified as $U_{nit} = bx_{nit} + \eta_n x_{nit} + \varepsilon_{nit}$, where the stochastic portion of utility is now correlated across alternatives through the attributes in the model. The conditional probability that individual n will choose alternative j at time period t is given by:

$$F_{nit}(\beta_n) = \frac{e^{(\beta_n x_{nit})}}{\sum_{j=1}^{J} e^{(\beta_n x_{njt})}} \tag{9}$$

The conditional probability of assigning a ranking for the full choice set of alternatives $j^1, j^2, j^3, \ldots, j^{J-1}$ can be calculated as the product of the conditional probability for all choice decisions of the exploded data:

$$\Pr_{nt}\left(j^1, j^2, j^3, \ldots, j^{J-1} \,|\theta^*\right) = \prod_{j=1}^{J-1} \frac{e^{(\beta_n x_{nit})}}{\sum_{j=1}^{J} e^{(\beta_n x_{njt})}} \tag{10}$$

Integration of the conditional probability over all possible values of β_n where the parameters θ^* define the distribution of β_n gives the unconditional probability that a given alternative j is chosen yields:

$$\Pr_{nt}\left(j_1, j_2, j_3, \ldots, j_{J-1} \,|\theta^*\right) = int \prod_{j=1}^{J-1} \frac{e^{(\beta_n x_{nit})}}{\sum_{j=1}^{J} e^{(\beta_{njt})}} f\left(\beta_n \,|\theta^*\right) \mathrm{d}\beta_n \tag{11}$$

Thus, for estimating θ^*, which are the parameters defining the distribution of coefficients β_n, the log-likelihood function to be maximized is:

$$\mathrm{LL}(\theta) = \sum_{n=1}^{N} \ln int \prod_{j=1}^{J-1} \frac{e^{(\beta_n x_{nit})}}{\sum_{j=1}^{J} e^{(\beta_{njt})}} f\left(\beta_n \,|\theta^*\right) \mathrm{d}\beta_n \tag{12}$$

The integral to be maximized in (12) has no closed-form solution; therefore, a numerical approximation is done using simulated maximum likelihood estimation (Train 2009; Greene 2012).

2.5.1 Experimental procedures and data

A total of 202 individuals (nonstudents) were recruited for the experiment using newspapers and internet ads. Participants were assigned to one of eight sessions based on schedule availability and according to their age and gender to reflect demographic characteristics of US shoppers (Carpenter and Moore 2006). Table 1 shows demographic summary statistics of participants. The experiment consisted of two parts: a ranking procedure and an auction bid. In the first part, participants were asked to rank the relative desirability of seven fruit products plus an 'opt-out' alternative for 'no-purchase' as described in the proposed ranking procedure section. The fruit products included in the experiment were: (1) California wonderful fresh pomegranate (the predominant variety in the market), (2) Texas red fresh pomegranate, (3) Texas Salavatski fresh pomegranate, (4) ready-to-eat California pomegranate arils, (5) ready-to-eat Texas pomegranate arils, (6) pomegranate juice, and (7) a pineapple as a control fruit. Pomegranates were selected because they are relatively new in the market and also based on funding availability. None of the Texas products were available in retail markets.

Table 1 Demographic Characteristics of participants

Variable	Category	Mean	SD	%
Age (years)		42.84	17.51	
	18–29			34.8
	30–49			26.4
	50–69			21.9
	70 or more			16.9
Education	High school diploma or less			11.4
	4-year college degree or less			60.7
	Graduate courses or more			27.9
Household size (number of individuals)		2.24	1.15	
	1			28.9
	2			39.3
	3			15.9
	4 or more			15.9
Gender	Female			68.7
	Male			31.3
Marital status	Married			54.2
	Not married			45.8
Household income ($/year)		53,693	36,973	
	<$30,000			31.2
	$30,000–$49,999			22.6
	$50,000–$79,999			24.6
	$80,000 or more			21.6
Primary shopper				88.0

The use of an unfamiliar product allows for testing how familiarity with the products affects choices by having a treatment where participants taste each product. Each session ranged from 20 to 30 participants. Subjects received a $35 compensation fee for their participation paid in cash, minus any purchases they made during the experiment. A combination of an 11th price sealed bid (Vickrey 1961) nonhypothetical auction and a nonhypothetical ranking procedure were used to elicit consumer preferences. The objective of the auction was to determine a market price for the products. In an 11th price auction, the 11th highest price is the market price. Hence the market price was close to the median bid for each session. Based on the ranking procedure, some participants would have to buy a product at the market price. This was done to incentivize participants to provide their true rankings. The nonhypothetical ranking mechanism was modeled following Lusk et al. (2008). The probability that a product j with ranking r was randomly selected as binding is given by: $\Pr_{lr} = \frac{J+1-r_1}{\sum_{j=1}^{J} j} \times 100$. The probability for a product being selected is higher for the most preferred alternatives

and lower for the least preferred alternatives, with at least one product being selected. If product j is selected, then participants would pay the market price based on the auction procedure, making the mechanism nonhypothetical.

Prior to eliciting ranking and bidding preferences, participants received extensive instructions about the mechanism and were explained that they would have to pay for any purchases they made. In addition, there were two practice rounds with soft drinks and snacks to train participants with the procedures. For the practice rounds, market prices were posted to ensure participants understood the procedures, but prices were not posted after each treatment to avoid bid affiliation (Corrigan and Rousu 2006). Subjects participated in four rounds of rankings and auctions for all the fruit products. The first round was a baseline round and no information was provided about the products. Following the baseline, participants were provided with three treatment rounds and were asked to submit their rankings and bids after each treatment. The treatments included two information treatments for nutritional information and anti-cancer properties of the fruit and a tasting treatment. For the tasting treatment, participants had the opportunity to taste a small sample (about two oz.) of each product. The tasting was voluntary, but none of the participants refused to taste any of the products. The tasting treatment is used to evaluate the effect of familiarity and preferences after purchase in a real life setting, where a consumer goes to a store, evaluates competing products, decides to purchase some of the products, and updates his/her preferences after tasting the purchased products. After conducting the treatments, the market price and buyers were announced. Only one round and one product were randomly selected as binding.

3 Results and discussion

The parameter estimates for the ROL model of the exploded data for the full choice set including all available rankings, and for the restricted choice consideration set with alternatives ranked above the opt-out choice are presented in Table 2. The results show the parameter estimates with different number of rankings, including the best alternative (rank = 1), partial rankings (rank = 2–rank = 5), and the full ranking (rank = 6). As the number of rankings increases, the number of independent observations also increases. For the full choice set ROL model the number of observations increased from 2782 (with rank = 1) to 10,650 (with rank = 6). The model with the restricted choice set, hereafter the heterogeneous ROL (hROL) had less observations since the choice set is restricted to those alternatives ranked higher than the opt-out alternative (2552 for rank = 1–8401 for rank = 6). The utility parameters include product attributes and the tasting information treatment. The tasting treatment captures changes in the choice variable as participants became more familiar with the products. Price is usually included in hypothetical discrete choice experiments; it was not included as an explanatory variable because the market price in the nonhypothetical mechanism was determined in the auction procedure, and it is highly correlated with the ranking.

Using more rankings increases the sample size leading to a reduction in the variance of the parameters. This is observed in Table 2 with a reduction of the standard errors of all parameters as the number of rankings increases. When examining the predictive power of the models, the proposed mechanism improves the prediction of

Table 2 Rank-ordered logit parameter estimates of the exploded data for the full and restricted choice set

	Full choice set[a]						Restricted choice set[a]						Best–worst
	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	
Variety													
Texas red	−1.2936**	−0.2951	0.0773	−0.0224	−0.1792	0.0149	−1.0654*	−0.2932	0.0295	−0.1728	−0.2438	−0.1624	0.1426
	(0.555)	(0.284)	(0.189)	(0.152)	(0.124)	(0.107)	(0.566)	(0.302)	(0.212)	(0.181)	(0.153)	(0.137)	(0.183)
Texas Salavatski	−0.6908**	0.0228	0.2365**	0.0968	0.0757	0.1248	−0.6131*	−0.1084	0.1423	0.0410	−0.0375	−0.0153	0.1312
	(0.327)	(0.178)	(0.121)	(0.100)	(0.086)	(0.081)	(0.344)	(0.196)	(0.140)	(0.118)	(0.105)	(0.100)	(0.142)
Product form													
Ready-to-eat (RTE)	−0.0782	0.4015**	0.6516***	0.6995***	0.4998***	0.5254***	0.0541	0.3674*	0.5258***	0.5464***	0.4123***	0.4272***	0.3686**
	(0.309)	(0.182)	(0.126)	(0.104)	(0.093)	(0.089)	(0.329)	(0.199)	(0.144)	(0.121)	(0.112)	(0.107)	(0.144)
Juice	0.9867***	1.5847***	1.4197***	1.1945***	0.8912***	0.7911***	0.9765***	1.3404***	1.1623***	0.9413***	0.6828***	0.551***	0.6400***
	(0.290)	(0.194)	(0.158)	(0.136)	(0.123)	(0.119)	(0.320)	(0.211)	(0.176)	(0.154)	(0.142)	(0.139)	(0.200)
Pineapple (control)	2.1309***	2.3493***	2.1463***	1.8662***	1.5088***	1.4439***	2.1869***	2.2167***	2.0030***	1.7583***	1.4722***	1.4276***	2.1495***
	(0.259)	(0.197)	(0.160)	(0.140)	(0.128)	(0.124)	(0.283)	(0.208)	(0.173)	(0.153)	(0.142)	(0.138)	(0.183)
Information treatment interactions													
Tasting × Texas red	1.9940***	0.7324**	0.4475*	0.4289**	0.5981***	0.4303***	1.5605**	0.7165*	0.4516*	0.5535**	0.6332***	0.525***	0.7056**
	(0.673)	(0.362)	(0.247)	(0.201)	(0.170)	(0.151)	(0.686)	(0.381)	(0.270)	(0.230)	(0.199)	(0.182)	(0.274)
Tasting × Texas Salavatski	0.5810	0.0044	−0.0066	0.1761	0.2402*	0.1707	0.3254	0.1212	0.0018	0.2157	0.2892**	0.2196	0.2204
	(0.461)	(0.245)	(0.170)	(0.141)	(0.123)	(0.116)	(0.485)	(0.264)	(0.191)	(0.161)	(0.144)	(0.136)	(0.206)
Tasting × RTE	0.3968	0.0647	−0.2757	−0.4474***	−0.2694**	−0.2932**	0.1237	0.0406	−0.2522	−0.4355***	−0.2735*	−0.2808**	−0.0090
	(0.451)	(0.251)	(0.175)	(0.144)	(0.129)	(0.122)	(0.473)	(0.269)	(0.195)	(0.163)	(0.149)	(0.142)	(0.208)
Tasting × juice	0.0531	−0.5169*	−0.7792***	−0.8917***	−0.7668***	−0.9487***	−0.1937	−0.4482	−0.7504***	−0.8160***	−0.7461***	−0.9057***	−1.0800***
	(0.461)	(0.277)	(0.225)	(0.195)	(0.177)	(0.169)	(0.486)	(0.299)	(0.248)	(0.217)	(0.199)	(0.193)	(0.268)
Tasting × pineapple	0.5522	−0.0979	−0.3036	−0.3412*	−0.2001	−0.2783	0.3793	0.0337	−0.1945	−0.2348	−0.1570	−0.2551	0.1451
	(0.403)	(0.274)	(0.221)	(0.194)	(0.180)	(0.173)	(0.415)	(0.287)	(0.235)	(0.208)	(0.194)	(0.187)	(0.263)

Table 2 continued

	Full choice set[a]						Restricted choice set[a]						Best–worst
	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	
NOBS[b]	2782	5152	7124	8698	9872	10650	2552	4505	5992	7129	7909	8401	5152
Tjur R^{2}[c]	0.297	0.172	0.089	0.072	0.062	0.088	0.309	0.170	0.091	0.074	0.065	0.092	0.137
Pseudo R^2	0.323	0.188	0.102	0.074	0.051	0.044	0.330	0.185	0.103	0.077	0.057	0.053	0.138
Compliance test[d]	0.600	0.516	0.422	0.397	0.394	0.432	0.611	0.508	0.403	0.397	0.387	0.413	0.470
LR-test stability of preferences[e]	20.23***	24.52***	35.72***	40.08***	42.69***	57.26***	–	–	–	–	–	–	221.75***

[a] *, **, and *** Significance at the 0.10, 0.05, and 0.01 levels, respectively (SE)

[b] Number of observations of the 'exploded' data

[c] Tjur Coefficient of Discrimination Goodness of fit measure, independent of the number of observations.

[d] The compliance test shows the proportion of the chosen alternatives with the highest predicted choice probability.

[e] LR test of stability of preferences. Three separate models are compared: ranking stage 1 (below opt-out), ranking stage 2 (above opt-out) and pooled data from the two ranking stages. The null hypothesis is that preferences are stable across the ranking stages

the models over the conventional ranking method. The models which include only the best ranked alternative (rank = 1) have the highest predictive power. A small increase in the prediction of the best alternative is observed from the conventional ranking method to the proposed mechanism with Tjur's R^2 going from 0.297 to 0.309 and compliance from 0.600 to 0.611. The predictive power of the models decreases as the number of rankings increases with one important exception. Notably, including the worst ranked alternative increases the explanatory power of both ranking methods compared to models using middle ranks (ranks 4–5).

This issue is further explored by estimating (case 1) best–worst models of the exploded data as described in Eq. (8). The BW models use the same number of observations as the best two ranking model. The Tjur's R^2 (0.137) are lower for the BW ranking than for the models with the best two alternatives. The compliance test of the proportion of cases in which the chosen alternative had the highest predicted choice probabilities shows similar results. The model with the highest overall correct prediction of the most preferred alternative is the partial ranking of the top (best) alternative only (rank = 1) for the restricted choice set (0.611). If two rankings are used, then using the best and second-best alternatives yields higher correct predictions compared to using exploded BW rankings. These results were robust across all logit specifications estimated for the data set. This provides evidence that using exploded BW rankings as a data augmentation tool leads to inferior predictive power compared to using the best and second-best alternatives. The results are in agreement with those of Hensher and Ho (2014) and as they point out raise some concerns about the behavioral validity of exploded BW models when used as a data augmentation tool.

The models in Table 2 assume constant error variances across the two ranking stages. Table 3 presents the results for the exploded rankings allowing for scale differences in the error variance. The scale parameter for alternatives ranked below and above the opt-out choice was not statistically significant for rank = 1–rank = 5, suggesting preferences are symmetrical for alternatives ranked above and below the opt-out choice for ranks = 1–5. Scale differences were found for the full ranking model (rank = 6). The scale was higher for alternatives ranked higher than no-product (lower variance) in the full ranking model.

A scale parameter of the rankings as a function of age and household size was also estimated. Results suggest that when middle rankings are used (rank = 2–rank = 5) older respondents have a larger scale (lower variance) for products ranked below no-product. Larger households had a smaller scale (larger variance) for the models which include the less preferred alternatives (rank = 4–rank = 6). The scale parameter of the BW model was statistically significant and lower (higher variance) for the worst alternative. When scale was made a function of age and household size, the BW model had a lower scale (higher variance) for larger households.

A structural change in the parameters to test for stability across choices above and below the defined consideration set was conducted by estimating separate models of the pooled data, and above and below the opt-out choice. A likelihood ratio test was calculated as $\text{LR} = -2[L_{r1+r2} - (L_{r1} + L_{r2})]$ where $r1$ is the data from the ranking stage 1, $r2$ is the data from the ranking stage 2, and $r1 + r2$ is the pooled data from the two ranking stages. As the number of rankings increases, the number of alternatives preferred to the opt-out choice increases (230 for rank = 1–2249 for rank = 6).

Table 3 Heteroscedastic error variance models of the exploded data for different rankings

	Scale ranked below no-product[a]						Scale age and household size[a]						Parameter (SE)[a]	
	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Best–worst I	Best–worst II
Variety														
Texas red	−1.1814**	−0.3346	0.0950	−0.0178	−0.1694	−0.0191	−1.3234**	−0.2031	0.1168	−0.0097	−0.1863	0.0208	−1.2934**	0.1748
	(0.557)	(0.327)	(0.214)	(0.157)	(0.114)	(0.082)	(0.635)	(0.251)	(0.171)	(0.148)	(0.134)	(0.121)	(0.559)	(0.226)
Texas Salavatski	−0.6333**	0.0475	0.2789**	0.1017	0.0600	0.0675	−0.7191**	0.0353	0.2250**	0.0930	0.0852	0.1368	−0.6907**	0.1559
	(0.323)	(0.205)	(0.140)	(0.104)	(0.081)	(0.066)	(0.360)	(0.157)	(0.112)	(0.097)	(0.090)	(0.094)	(0.327)	(0.177)
Product form														
Ready-to-eat (RTE)	−0.0618	0.4617**	0.7508***	0.7252***	0.4469***	0.3666***	−0.0662	0.3699**	0.6184***	.7190***	0.5700***	0.6474***	−0.0781	0.4937***
	(0.287)	(0.214)	(0.163)	(0.136)	(0.108)	(0.105)	(0.331)	(0.165)	(0.140)	(0.143)	(0.134)	(0.146)	(0.311)	(0.188)
Juice	0.9110***	1.8386***	1.6502***	1.2425***	0.7848***	0.5252***	1.0661***	1.4148***	1.3254***	1.1981***	0.9831***	0.9526***	0.9868***	0.7771***
	(0.309)	(0.297)	(0.257)	(0.210)	(0.174)	(0.156)	(0.327)	(0.236)	(0.224)	(0.213)	(0.204)	(0.205)	(0.297)	(0.268)
Pineapple (control)	1.9667***	2.7181***	2.4822***	1.9361***	1.3475***	1.021***	2.2509***	2.0806***	1.9690***	1.8371***	1.6127***	1.6771***	2.131***	2.6808***
	(0.408)	(0.367)	(0.326)	(0.270)	(0.235)	(0.233)	(0.396)	(0.305)	(0.299)	(0.291)	(0.288)	(0.303)	(0.267)	(0.422)
Information treatment interactions														
Tasting × Texas red	1.8085**	0.8417**	0.5158*	0.4400**	0.5418***	0.3189**	2.0764***	0.5982*	0.3962*	0.4061**	0.6197***	0.4809***	1.9938***	0.8948**
	(0.718)	(0.425)	(0.287)	(0.211)	(0.170)	(0.126)	(0.796)	(0.327)	(0.229)	(0.202)	(0.199)	(0.186)	(0.684)	(0.365)
Tasting × Texas Salavatski	0.5187	−0.013	0.0013	0.1820	0.2167*	0.1234	0.6367	−0.0044	0.0041	0.1719	0.2510*	0.2034	0.5810	0.2840
	(0.442)	(0.282)	(0.194)	(0.147)	(0.116)	(0.088)	(0.494)	(0.215)	(0.153)	(0.137)	(0.133)	(0.134)	(0.461)	(0.259)
Tasting × RTE	0.3431	0.0881	−0.3004	−0.4585***	−0.2499**	−0.2200**	0.3786	0.0300	−0.2637	−0.4626***	−0.3052**	−0.3523**	0.3967	−0.0106
	(0.428)	(0.289)	(0.200)	(0.153)	(0.119)	(0.098)	(0.484)	(0.222)	(0.161)	(0.152)	(0.142)	(0.149)	(0.452)	(0.258)
Tasting × juice	0.0239	−0.5756*	−0.8708***	−0.9217***	−0.6900***	−0.6699***	−0.0064	−0.4917**	−0.7331***	−0.9055***	−0.8460***	−1.1314***	0.0529	−1.3838***
	(0.428)	(0.321)	(0.266)	(0.224)	(0.187)	(0.187)	(0.485)	(0.250)	(0.227)	(0.230)	(0.235)	(0.275)	(0.476)	(0.385)
Tasting × pineapple	0.4833	−0.1027	−0.3347	−0.3509*	−0.1854	−0.2115	0.5384	−0.1342	−0.3168	−0.3833**	−0.2550	−0.3522*	0.5521	0.1711
	(0.393)	(0.316)	(0.256)	(0.204)	(0.162)	(0.129)	(0.436)	(0.241)	(0.204)	(0.195)	(0.191)	(0.203)	(0.421)	(0.328)

Table 3 continued

	Scale ranked below no-product[a]						Scale age and household size[a]						Parameter (SE)[a]	
	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Best–worst I	Best–worst II
Scale ranked below no product	0.0910	−0.1654	−0.1696	−0.0433	0.1320	0.3958*	–	–	–	–	–	–	–	–
	(0.189)	(0.124)	(0.132)	(0.140)	(0.173)	(0.224)	–	–	–	–	–	–	–	–
Scale worst	–	–	–	–	–	–	–	–	–	–	–	–	−17.5346***	–
	–	–	–	–	–	–	–	–	–	–	–	–	(0.292)	–
Scale age	–	–	–	–	–	–	0.0182	0.0508**	0.0592**	0.0575**	0.0532*	0.0375	–	−0.0093
	–	–	–	–	–	–	(0.028)	(0.025)	(0.027)	(0.028)	(0.032)	(0.033)	–	(0.029)
Scale household size	–	–	–	–	–	–	−0.0486	−0.0261	−0.0544	−0.0845**	−0.1136**	−0.1248***	–	−0.0793*
	–	–	–	–	–	–	(0.040)	(0.035)	(0.038)	(0.040)	(0.046)	(0.047)	–	(0.041)
LM heteroscedasticity[b]	0.628	0.176	0.191	0.756	0.441	0.059	0.388	0.088	0.029	0.010	0.007	0.011	0.000	0.137
NOBS[c]	2782	5152	7124	8698	9872	10650	2755	5102	7055	8614	9777	10549	5155	5105
Tjur R^2	0.273	0.200	0.105	0.075	0.058	0.078	0.311	0.146	0.079	0.070	0.065	0.094	0.141	0.179
LL(B)	−524.992	−1205.088	−1904.583	−2469.710	−2938.614	−3216.712	−518.501	−1194.543	−1883.831	−2439.948	−2903.813	−3181.282	−1234.726	−1263.222
LL(0)	−775.436	−1485.047	−2121.210	−2667.306	−3096.916	−3366.550	−775.436	−1485.047	−2121.210	−2667.306	−3096.916	−3366.550	−1485.047	−1485.047
Pseudo R^2	0.323	0.189	0.102	0.074	0.051	0.045	0.331	0.196	0.112	0.085	0.062	0.055	0.169	0.149
Compliance test[d]	0.600	0.516	0.422	0.397	0.394	0.422	0.600	0.516	0.422	0.397	0.394	0.432	0.357	0.474

[a] *, **, and *** Significance at the 0.10, 0.05, and 0.01 levels, respectively (SE)

[b] LM test for heteroscedastic error variance. Reported p value

[c] Number of observations of the 'exploded' data

[d] The compliance test shows the proportion of the chosen alternatives with the highest predicted choice probability

Intuitively, respondents are more likely to select the opt-out choice for less preferred alternatives. The pooling hypothesis is rejected for all rankings. Hence, preferences are not stable across the choices made below and above the opt-out choice. It can be concluded that the decision rules that determine preferences are different across the relevant consideration set and the alternatives outside of the consideration set. The statistical significance of the test for the stability of the parameters increases as the number of rankings increase (i.e., using less preferred alternatives provides more evidence of non-stable parameters across the ranking stages). A stability of parameters test was also conducted following (Allison and Christakis 1994). The approach involves estimating stage dependent covariates with product terms for the parameters across stages. This approach resulted in the rejection of stability of preferences for all rankings across the two stages (below and above the opt-out choice). These results are available in an appendix. A test of the stability of parameters for the exploded BW ranking method was conducted and revealed that the parameters for the best alternative differ structurally from the parameters of the worst ranked alternative ($\chi^2 = 221.75$). What about the stability of parameters of going from the best alternative to pooling the second-best alternatives? Even going from the best ranked alternative to the second-best best ranked alternatives resulted in nonstable parameter estimates ($\chi^2 = 79.84$). Although using the proposed restricted choice set reduced the statistical significance of the stability test, the results were nevertheless the same; the parameters were not stable across the best-rank and the second-best rankings ($\chi^2 = 70.38$). The results challenge the common practice of pooling data from different rankings. In this case, only the best ranked alternative can be used because preferences are not stable across any of the ranking stages. The model with the highest predictive power comes from the proposed ranking method using only the best ranking alternative (Tjur = 0.309, compliance = 0.611). Using any additional rankings other than the best alternative would be inappropriate since it is clear that the decision rules governing preferences for the best alternative are different than all other alternatives. Nevertheless, if more than one ranking is used to model preferences, then better predictions are achieved by using the top two best ranked alternatives as supposed to the exploded BW.

In order to capture individual heterogeneity of preferences, a random parameter ROL model was estimated. The Random Parameter ROL model of the exploded data was estimated in STATA 13.0 based on the procedure in Hole (2007) using 500 Halton draws. The estimated parameters for the RPROL are presented in Table 4. Table 5 presents the parameter estimates for the RPROL of the proposed ranking mechanism restricted to observations ranked higher than the no-product option.[3]

Although the pattern is not always monotonic, a similar trend as in the constant parameter models is observed with a reduction of the standard errors of the parameter estimates as the number of rankings increases. The standard deviations of the random parameters are highly significant showing the presence of unobserved taste heterogeneity in the choice variable. The predictive power of the traditional ranking method and the proposed heterogeneous ranking method show a similar pattern as the

[3] The mixlogit model for the top rank was not identified since there was only one decision and no panel structure; convergence was achieved with a low number of draws masking an identification problem as described by Chiou and Walker (2007).

Table 4 Random parameter rank-ordered logit estimates for the traditional ranking method with homogeneous rankings

	Means of random parameters[a]						Standard deviations of random parameters[a]					
	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Best–worst	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Best–worst
Variety												
Texas red	−0.4475	−0.0918	−0.1171	−0.2533*	0.0265	0.1923	−0.3601	−0.2873	−0.3071*	−0.2279	−0.1328	0.5375
	(0.309)	(0.211)	(0.169)	(0.138)	(0.118)	(0.224)	(0.821)	(0.240)	(0.184)	(0.162)	(0.159)	(0.343)
Texas Salavatski	−0.088	0.2265	0.1617	0.1373	0.2018**	0.1610	0.5577*	0.4702**	0.0284	0.1107	−0.3093**	−0.2526
	(0.204)	(0.138)	(0.112)	(0.098)	(0.094)	(0.168)	(0.336)	(0.211)	(0.153)	(0.195)	(0.147)	(0.471)
Product form												
Ready-to-eat (RTE)	0.4483	1.2399***	1.3600***	1.1167***	1.0468***	0.7756***	2.8231***	2.7826***	2.857***	3.0303***	2.4796***	1.4447***
	(0.358)	(0.258)	(0.267)	(0.222)	(0.188)	(0.231)	(0.485)	(0.305)	(0.253)	(0.271)	(0.177)	(0.291)
Juice	3.4064***	3.9468***	3.1580***	2.1470***	2.5397***	1.5578***	12.1186***	9.1858***	5.9564***	6.0993***	6.5065***	3.1815***
	(1.061)	(0.776)	(0.605)	(0.446)	(0.378)	(0.373)	(3.186)	(1.215)	(0.594)	(0.542)	(0.515)	(0.405)
Pineapple (control)	11.3145***	11.8295***	7.4924***	7.5563***	8.0087***	4.9465***	20.3953***	16.5098***	8.7470***	9.7278***	10.631***	5.1276***
	(2.080)	(1.413)	(0.694)	(0.717)	(0.659)	(0.649)	(5.496)	(2.417)	(0.803)	(0.975)	(0.840)	(0.696)
Information treatment interactions												
Tasting × Texas red	0.3975	0.8378***	0.7868***	0.9986***	0.7231***	1.2151***	−2.4181**	−0.7801**	−0.3991	0.7234**	−0.0599	0.1019
	(0.731)	(0.286)	(0.231)	(0.208)	(0.176)	(0.348)	(0.969)	(0.358)	(0.412)	(0.326)	(0.293)	(0.636)
Tasting × Texas Salavatski	0.0338	0.0465	0.2498	0.4708***	0.3200**	0.3533	−0.3032	−0.4498	0.9955***	−1.1732***	−0.9081***	−0.3598
	(0.276)	(0.195)	(0.189)	(0.166)	(0.150)	(0.250)	(0.617)	(0.564)	(0.224)	(0.182)	(0.169)	(0.446)
Tasting × RTE	0.1940	−0.6560**	−0.7949***	−0.6687***	−0.4836***	−0.0590	−1.0843	0.6644*	−0.0012	1.028***	1.039***	0.3859
	(0.382)	(0.255)	(0.205)	(0.194)	(0.183)	(0.289)	(0.718)	(0.360)	(0.998)	(0.293)	(0.212)	(0.434)
Tasting × juice	−7.6671***	−4.9887***	−3.7848***	−2.8327***	−3.1686***	−1.7836***	−12.6655***	8.801***	8.5325***	5.2538***	5.2889***	3.5781***
	(2.542)	(0.984)	(0.677)	(0.520)	(0.487)	(0.479)	(3.223)	(1.251)	(1.063)	(0.533)	(0.483)	(0.678)
Tasting × pineapple	−0.1767	0.1659	−0.6320	−0.6608*	−0.4449	0.5538	−9.0778***	9.8834***	−4.3008***	−0.7133*	2.9018***	2.807***
	(0.938)	(1.028)	(0.498)	(0.344)	(0.408)	(0.564)	(2.505)	(1.656)	(0.728)	(0.400)	(0.412)	(0.849)

Table 4 continued

	Means of random parameters[a]						Standard Deviations of random parameters[a]					
	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Best–worst	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Best–worst
NOBS[b]	5152	7124	8698	9872	10650	5152						
Tjur R^2[c]	0.209	0.156	0.132	0.120	0.140	0.183						
Pseudo R^2	0.139	0.188	0.182	0.183	0.185	0.134						
Compliance test[d]	0.513	0.423	0.409	0.389	0.427	0.461						

[a] *, **, and *** Significance at the 0.10, 0.05, and 0.01 levels, respectively (SE)

[b] Number of observations of the 'exploded' data

[c] Tjur coefficient of discrimination goodness of fit measure, independent of the number of observations

[d] The compliance test shows the proportion of the chosen alternatives with the highest predicted choice probability

Table 5 Random parameter rank-ordered logit estimates for the proposed heterogeneous ranking method

	Means of random parameters[a]					Standard deviations of random parameters[a]				
	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6
Variety										
Texas red	−0.6647	−0.1488	−0.2794	−0.4595**	−0.2966*	1.018*	−0.3879	−0.3425	−0.5400**	−0.1582
	(0.417)	(0.241)	(0.204)	(0.187)	(0.157)	(0.541)	(0.305)	(0.220)	(0.225)	(0.236)
Texas Salavatski	−0.2340	0.1216	0.0307	−0.0472	−0.0369	0.0165	−0.3917	−0.3680**	0.3466**	−0.572**
	(0.212)	(0.160)	(0.140)	(0.124)	(0.125)	(0.370)	(0.279)	(0.184)	(0.173)	(0.229)
Product form										
Ready-to-eat (RTE)	0.1881	0.4806	0.8419***	0.9077***	0.9753***	3.3315***	3.2663***	3.6571***	3.1552***	2.6163***
	(0.425)	(0.301)	(0.297)	(0.231)	(0.335)	(0.647)	(0.369)	(0.404)	(0.284)	(0.267)
Juice	2.1263	3.4454***	3.1750***	1.9540***	2.2544***	10.5914***	10.3007***	6.5505***	6.3061***	6.5803***
	(1.361)	(0.644)	(0.591)	(0.442)	(0.645)	(2.446)	(1.420)	(0.784)	(0.565)	(0.558)
Pineapple (control)	10.1309***	10.2896***	7.1067***	6.7448***	6.6321***	16.3197***	12.7785***	8.8691***	6.8849***	7.0036***
	(2.004)	(1.247)	(0.779)	(0.619)	(0.919)	(4.186)	(1.554)	(0.985)	(0.556)	(0.570)
Information treatment interactions										
Tasting × Texas red	1.0331**	0.8311**	0.9219***	1.1304***	1.0026***	−1.2816	−0.8510**	0.5994*	−1.2211***	−0.6758
	(0.509)	(0.326)	(0.270)	(0.262)	(0.235)	(0.780)	(0.420)	(0.349)	(0.289)	(0.519)
Tasting × Texas Salavatski	0.1243	0.0673	0.2863	0.4852**	0.3990**	−0.9857	0.9574***	−1.094***	1.1245***	−0.7759***
	(0.323)	(0.232)	(0.209)	(0.191)	(0.180)	(0.664)	(0.355)	(0.283)	(0.223)	(0.276)
Tasting × RTE	−0.0065	−0.8535**	−0.7525***	−0.5986***	−0.5671**	0.7986	1.9076***	1.1770***	0.5772**	0.5789**
	(0.434)	(0.336)	(0.265)	(0.228)	(0.221)	(0.777)	(0.378)	(0.268)	(0.275)	(0.274)
Tasting × juice	−5.6667***	−5.6814***	−4.6305***	−3.0319***	−3.8127***	−7.3629***	7.5951***	5.3805***	−4.2077***	−5.5799***
	(2.052)	(1.079)	(1.022)	(0.547)	(0.584)	(1.872)	(1.376)	(0.779)	(0.490)	(0.628)
Tasting × pineapple	0.6769	−0.3134	0.4959	0.0104	−0.2093	−10.6033***	−5.3063***	6.2258***	4.0715***	4.4880***
	(1.372)	(0.862)	(0.678)	(0.450)	(0.537)	(2.908)	(1.027)	(0.775)	(0.528)	(0.598)

Table 5 continued

	Means of random parameters[a]					Standard deviations of random parameters[a]				
	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6
NOBS[b]	4505	5992	7129	7909	8401					
Tjur R^{2}[c]	0.212	0.154	0.133	0.124	0.141					
Pseudo R^2	0.155	0.195	0.190	0.190	0.192					
Compliance test[d]	0.508	0.411	0.396	0.376	0.408					

[a] *, **, and *** Significance at the 0.10, 0.05, and 0.01 levels, respectively (SE)
[b] Number of observations of the 'exploded' data
[c] Tjur coefficient of discrimination goodness of fit measure, independent of the number of observations
[d] The compliance test shows the proportion of the chosen alternatives with the highest predicted choice probability

constant parameter models, with decreases for the middle ranks and an increase in the prediction accuracy when using the worst alternative in the full ranking (rank = 6) compared to middle ranks (ranks 4–5). The exploded BW ranking showed similar results as before with lower Tjur's R^2 than the models with the best two alternatives. The compliance test also showed higher predictions of the chosen alternative using the best and second-best alternatives compared to the BW. If data augmentation is used and individual preference heterogeneity is incorporated in a random parameter framework, then using the top two (best) rankings results in better predictions than the BW method. The stability of the parameters hypothesis was tested using Allison and Christakis (1994) framework, and it was also rejected for all partial rankings (rank = 2–rank = 5) and the full rankings (rank = 6). Results of these tests are also available in "Appendix" section.

The proposed ranking elicitation mechanism robustly provided the best prediction accuracy compared to conventional full and partial ranking methods including the exploded BW method. There is supporting evidence for relaxing the assumption that respondents know their utility and are able to rank *the same* number of options. The compliance test also showed that if the models are used to predict the chosen alternative, using the best and second-best alternatives always have higher compliance of predicting the chosen alternative compared to exploded BW models. The results also provide robust evidence of differences in error variance scale and nonstable underlying utility for the alternatives ranked above and below the no-product option and for the best and worst ranked alternatives in exploded BW methods. Differences in preferences across ranking stages, including exploded BW types, may be due to respondents not knowing their utility for some of the alternatives, general misconceptions of the elicitation mechanism (Cason and Plott 2014), or complexity, tediousness and high cognitive costs of the ranking task itself. The results stand as a warning about equating ranking choices to true underlying utility preferences across different ranking elicitation stages or mechanisms without properly testing for symmetry and stability of the underlying preferences.

4 Summary and conclusions

The use of surveys to elicit consumer preferences using ranking mechanisms is very common in the literature. When full rankings are available, additional preference information can be analyzed by including more than just the most preferred alternative. Data collection is expensive, and additional rankings are usually elicited as a data augmentation tool to reduce the number of participants in surveys or economic experiments or to reduce the number of tasks a participant faces. The ranked-ordered logit model allows for the analysis of additional ranking data by using the full set of available alternatives in the choice set. All of the ranking elicitation mechanisms in the literature assume that respondents know their utility for and are able to rank *the same* number of choices. However, there are some reasons why using the full set of alternatives may not be appropriate. This article introduced a simple survey elicitation mechanism that allows for heterogeneity in the number of rankings for each participant, thus relaxing the assumption that respondent know the utility and are able

to rank *the same* number of alternatives. The method consists of a two-step process where respondents determine a consideration choice set composed of alternatives they would purchase. This is empirically tested by adding an additional alternative in the choice set for an opt-out 'no-product' option to serve as a natural endogenous break point for relevant alternatives. A heteroscedastic error variance model with a scale parameter for alternatives ranked below and above the opt-out choice was not statistically significant for rank=1 through rank = 5, suggesting preferences are symmetrical for alternatives ranked above and below the opt-out choice for ranks = 1–5. Scale differences were found for the full ranking model (rank = 6). The scale was higher for alternatives ranked higher than no-product (lower variance) in the full ranking model. More importantly, a test of the stability of parameters across ranking stages showed nonstable parameters for alternatives ranked below and above the 'no-product' alternative; statistical significance increased as the number of rankings increases. Utility preferences were not stable across best and worst ranked alternatives. Furthermore, even going from the best ranked alternative to the top two alternatives resulted in nonstable preferences. These results challenge the common practice of pooling data from different rankings into a single model without properly accounting for symmetry and testing for the stability of the underlying preferences across ranking stages. For the empirical dataset analyzed, only the best ranked alternative can be used because preferences are not stable across any of the ranking stages. The model with the highest predictive power comes from the proposed ranking method using the best ranking alternative. Using any additional rankings other than the best alternative would be inappropriate since it is clear that the decision rules governing preferences for the best alternative are different than all other alternatives.

Acknowledgements I received useful comments from Valentin Estevez, Ximing Wu, Jayson Lusk and two anonymous referees. I am grateful to Callie McAdams for help in the data collection stage.

Appendix

See Tables 6 and 7.

Table 6 Stage specific test of stability of the parameter estimates in a constant parameter model of the exploded data

	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6
Variety						
Texas red	−2.7978***	−2.0708***	−1.172***	−0.8979***	−0.9182***	−0.4696***
	(0.850)	(0.502)	(0.278)	(0.201)	(0.164)	(0.133)
Texas Salavatski	−1.6749***	−0.7753***	−0.3101*	−0.487***	−0.3213***	−0.1979*
	(0.569)	(0.269)	(0.162)	(0.134)	(0.111)	(0.104)
Product form						
Ready-to-eat (RTE)	−1.5347***	−0.642**	0.0603	0.2652**	0.036	0.0927
	(0.567)	(0.268)	(0.163)	(0.133)	(0.121)	(0.116)
Juice	−0.0864	0.8149***	0.803***	0.6218***	0.3857**	0.331**
	(0.408)	(0.245)	(0.212)	(0.188)	(0.173)	(0.167)
Pineapple (control)	0.3031	0.6853**	0.5843**	0.2858	−0.121	−0.303
	(0.398)	(0.314)	(0.283)	(0.260)	(0.249)	(0.241)
Information treatment interactions						
Tasting × Texas red	1.5861**	0.3343	0.0652	0.1183	0.3227*	0.2234
	(0.681)	(0.371)	(0.255)	(0.207)	(0.175)	(0.155)
Tasting × Texas Salavatski	0.3659	−0.1729	−0.1713	−0.0208	0.0793	0.0216
	(0.471)	(0.251)	(0.175)	(0.145)	(0.127)	(0.119)
Tasting × RTE	0.1101	−0.0979	−0.3947**	−0.5407***	−0.3913***	−0.4076***
	(0.469)	(0.260)	(0.180)	(0.147)	(0.132)	(0.125)
Tasting × juice	−0.2202	−0.6607**	−0.9282***	−1.0477***	−0.9254***	−1.1074***
	(0.467)	(0.280)	(0.230)	(0.200)	(0.181)	(0.174)
Tasting × pineapple	0.2403	−0.3291	−0.5416**	−0.5923***	−0.4638**	−0.565***
	(0.416)	(0.283)	(0.231)	(0.203)	(0.187)	(0.180)

Table 6 continued

	Rank = 1	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6
Stage dependent parameter estimates						
Ranked > NP × Texas red	2.3515***	2.6736***	2.1101***	1.5923***	1.3907***	0.9846***
	(0.784)	(0.492)	(0.276)	(0.205)	(0.172)	(0.146)
Ranked > NP × Texas Salavatski	1.4112**	1.1937***	0.9303***	1.0286***	0.7456***	0.6277***
	(0.584)	(0.275)	(0.171)	(0.143)	(0.122)	(0.115)
Ranked > NP × RTE	2.1045***	1.542***	0.96***	0.7086***	0.7979***	0.7564***
	(0.578)	(0.268)	(0.170)	(0.140)	(0.128)	(0.122)
Ranked > NP × juice	1.6719***	1.2073***	1.0652***	1.0147***	0.9347***	0.8661***
	(0.409)	(0.235)	(0.208)	(0.190)	(0.178)	(0.173)
Ranked > NP × pineapple	2.3873***	2.1696***	2.0931***	2.128***	2.2077***	2.3584***
	(0.396)	(0.304)	(0.278)	(0.260)	(0.251)	(0.245)
NOBS(c)	2782	5152	7124	8698	9872	10650
LL(B)	−491.311	−1138.433	−1825.529	−2381.195	−2845.461	−3127.585
LL(0)	−775.436	−1485.047	−2121.210	−2667.306	−3096.916	−3366.550
Pseudo R^2	0.366	0.233	0.139	0.107	0.081	0.071
Tjur R^2	0.333	0.210	0.123	0.102	0.089	0.112
Test stage param = 0	0.000	0.000	0.000	0.000	0.000	0.000

Table 7 Stage specific test of stability of the parameter estimates in a random parameter model of the exploded data

	Means of random parameters					Standard deviations of random parameters				
	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6
Variety										
Texas red	−2.3078***	−0.7484**	−0.5723**	−0.6888***	−0.1052	1.1326**	0.0169	−0.0789	0.3457**	−0.0443
	(0.787)	(0.321)	(0.236)	(0.195)	(0.156)	(0.448)	(0.257)	(0.249)	(0.176)	(0.165)
Texas Salavatski	−0.5548	0.0266	−0.1469	−0.0304	0.1074	0.9547***	0.7337***	0.121	−0.0611	0.5487***
	(0.373)	(0.214)	(0.166)	(0.138)	(0.140)	(0.272)	(0.169)	(0.151)	(0.209)	(0.117)
Product form										
Ready-to-eat (RTE)	−0.5382	0.6703**	0.5661*	0.3086	0.4683*	0.7915*	2.0094***	2.4654***	2.5098***	2.9986***
	(0.390)	(0.321)	(0.310)	(0.283)	(0.269)	(0.447)	(0.251)	(0.232)	(0.240)	(0.271)
Juice	2.3014**	1.5494	1.9136***	2.3375***	0.6075	7.5602***	9.3273***	4.2402***	3.2965***	4.8426***
	(1.174)	(1.140)	(0.540)	(0.586)	(0.589)	(1.931)	(1.694)	(0.543)	(0.398)	(0.404)
Pineapple (control)	−1.1754	6.4146***	2.8635*	1.0382	0.9501*	15.1851***	13.7132***	7.0042***	4.2822***	6.9674***
	(1.825)	(2.133)	(1.494)	(0.820)	(0.567)	(3.115)	(2.676)	(0.646)	(0.451)	(0.597)
Information treatment interactions										
Tasting × Texas red	1.0993**	0.6973**	0.6696***	0.8845***	0.7174***	1.7607**	0.4423	0.0475	0.4061	−0.1845
	(0.512)	(0.286)	(0.235)	(0.209)	(0.185)	(0.840)	(0.435)	(0.257)	(0.341)	(0.344)
Tasting × Texas Salavatski	0.011	−0.0468	0.1795	0.3673**	0.3557**	−1.1029*	−0.2124	−0.9625***	−1.0511***	−1.1062***
	(0.319)	(0.200)	(0.183)	(0.165)	(0.162)	(0.598)	(0.307)	(0.222)	(0.194)	(0.161)
Tasting × RTE	−0.2192	−0.7832***	−0.9802***	−0.6599***	−0.6767***	2.7694***	0.8784***	1.6402***	1.2565***	0.5725***
	(0.514)	(0.273)	(0.245)	(0.205)	(0.181)	(0.670)	(0.319)	(0.286)	(0.293)	(0.201)
Tasting × juice	−7.0531***	−6.1332***	−2.95***	−3.0626***	−3.7562***	−11.0655***	−7.9109***	−5.2194***	4.0612***	5.0688***
	(2.153)	(1.141)	(0.541)	(0.487)	(0.490)	(2.707)	(1.390)	(0.615)	(0.530)	(0.513)
Tasting × pineapple	−0.1727	−1.7867*	−1.2372**	−0.7597*	−1.1372***	10.8443***	−12.5723***	−6.4698***	1.7352**	2.4767***
	(1.177)	(0.975)	(0.552)	(0.390)	(0.411)	(2.473)	(2.936)	(0.802)	(0.707)	(0.364)

Table 7 continued

	Means of random parameters					Standard deviations of random parameters				
	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6	Rank = 2	Rank = 3	Rank = 4	Rank = 5	Rank = 6
Stage dependent parameter estimates										
Ranked > NP × Texas red	1.8002**	0.9971***	0.7816***	0.7401***	0.2475	0.8698**	0.39	0.1807	−0.2771	−0.5021***
	(0.760)	(0.336)	(0.255)	(0.218)	(0.189)	(0.440)	(0.265)	(0.211)	(0.226)	(0.187)
Ranked > NP × Texas Salavatski	0.5121	0.3168	0.4905***	0.2722	0.1828	−0.3934	−0.1918	0.355	−0.5117***	−0.1279
	(0.401)	(0.238)	(0.190)	(0.166)	(0.166)	(0.353)	(0.238)	(0.240)	(0.158)	(0.182)
Ranked > NP × RTE	1.4998***	0.9028**	1.4176***	1.284***	1.024***	5.2747***	2.7517***	2.0106***	1.41***	0.9866***
	(0.567)	(0.412)	(0.376)	(0.314)	(0.284)	(1.083)	(0.453)	(0.303)	(0.279)	(0.253)
Ranked > NP × juice	2.7992**	1.1925	1.3696**	2.7838***	1.8016**	7.788***	5.5401***	4.437***	5.8002***	4.701***
	(1.363)	(1.200)	(0.616)	(0.677)	(0.724)	(1.575)	(1.112)	(0.543)	(0.626)	(0.493)
Ranked > NP × pineapple	15.6635***	8.1082***	7.1323***	6.8994***	7.8313***	4.3309***	12.7505***	5.1121***	9.1725***	5.5691***
	(3.757)	(2.621)	(1.565)	(1.081)	(0.808)	(1.016)	(2.437)	(0.562)	(0.893)	(0.569)
NOBS(c)	5152	7124	8698	9872	10650					
LL(B)	−999.988	−1527.275	−1992.779	−2387.069	−2595.308					
LL(0)	−1138.433	−1825.529	−2391.195	−2845.461	−3127.585					
Pseudo R^2	0.122	0.163	0.167	0.161	0.170					
Tjur R^2	0.238	0.159	0.150	0.127	0.152					
Test stage param = 0	0.000	0.000	0.000	0.000	0.000					

References

Allison PD, Christakis NA (1994) Logit models for sets of ranked items. Sociol Methodol 24(1994):199–228

Auger P, Devinney TM, Louviere JJ (2007) Using best–worst scaling methodology to investigate consumer ethical beliefs across countries. J Bus Ethics 70(3):299–326. doi:10.1007/s10551-006-9112-7

Beggs S, Cardell S, Hausman J (1981) Assessing the potential demand for electric cars. J Econom 17(1):1–19. doi:10.1016/0304-4076(81)90056-7

Bronnenberg BJ, Vanhonacker WR (1996) Limited choice sets, local price response and implied measures of price competition. J Mark Res 33(2):163–173. doi:10.2307/3152144

Calfee J, Winston C, Stempski R (2001) Econometric issues in estimating consumer preferences from stated preference data: a case study of the value of automobile travel time. Rev Econ Stat 83(4):699–707. doi:10.1162/003465301753237777

Carpenter JM, Moore M (2006) Consumer demographics, store attributes, and retail format choice in the US grocery market. Int J Retail Distrib Manag 34(6):434–452. doi:10.1108/09590550610667038

Cason TN, Plott CR (2014) Misconceptions and game form recognition of the BDM method: challenges to theories of revealed preference and framing

Chapman RG, Staelin R (1982) Exploiting rank ordered choice set data within the stochastic utility model. J Mark Res 19(3):288–301. doi:10.2307/3151563

Chiang J, Chib S, Narasimhan C (1998) Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. J Econom 89(1–2):223–248. doi:10.1016/S0304-4076(98)00062-1

Chiou L, Walker JL (2007) Masking identification of discrete choice models under simulation methods. J Econom 141(2):683–703. doi:10.1016/j.jeconom.2006.10.012

Cohen E, Goodman S, Cohen E (2009) Applying best–worst scaling to wine marketing. Int J Wine Bus Res 21(1):8–23. doi:10.1108/17511060910948008

Corrigan JR, Rousu MC (2006) Posted prices and bid affiliation: evidence from experimental auctions. Am J Agric Econ 88(4):1078–1090. doi:10.1111/j.1467-8276.2006.00917.x

Finn A, Louviere JJ (1992) Determining the appropriate response to evidence of public concern: the case of food safety. J Public Policy Mark 11(2):12–25. doi:10.2307/30000270

Flynn TN, Louviere JJ, Peters TJ, Coast J (2007) Best–worst scaling: what it can do for health care research and how to do it. J Health Econ 26(1):171–189. doi:10.1016/j.jhealeco.2006.04.002

Fok D, Paap R, Van Dijk B (2012) A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. J Appl Econom 27(5):831–846. doi:10.1002/jae.1223

Giergiczny M, Hess S, Dekker T, Chintakayala PK (2013) Testing the consistency (or lack thereof) between choices in best–worst surveys. In: Third international choice modelling conference, Sydney

Greene WH (2012) Econometric analysis. Prentice Hall, Boston

Hausman JA, Ruud PA (1987) Specifying and testing econometric models for rank-ordered data. J Econom 34(1–2):83–104. doi:10.1016/0304-4076(87)90068-6

Hensher DA, Rose JM (2012) The influence of alternative acceptability, attribute thresholds and choice response certainty on automobile purchase preferences. J Transp Econ Policy (JTEP) 46(3):451–468

Hensher DA, Ho C (2014) Identifying a behaviourally relevant choice set from stated choice data. Transportation. doi:10.1007/s11116-014-9572-z

Hess S, Rose JM (2009) Allowing for intra-respondent variations in coefficients estimated on repeated choice data. Transp Res Part B Methodol 43(6):708–719. doi:10.1016/j.trb.2009.01.007

Hole AR (2007) Estimating mixed logit models using maximum simulated likelihood. Stata J 7(3):388–401

Koop G, Poirier DJ (1994) Rank-ordered logit models: an empirical analysis of Ontario voter preferences. J Appl Econom 9(4):369–388. doi:10.1002/jae.3950090406

Louviere J, Lings I, Islam T, Gudergan S, Flynn T (2013) An introduction to the application of (case 1) best–worst scaling in marketing research. Int J Res Mark 30(3):292–303. doi:10.1016/j.ijresmar.2012.10.002

Lusk JL, Fields D, Prevatt W (2008) An incentive compatible conjoint ranking mechanism. Am J Agric Econ 90(2):487–498

Marley AAJ, Pihlens D (2012) Models of best–worst choice and ranking among multiattribute options (profiles). J Math Psychol 56(1):24–34. doi:10.1016/j.jmp.2011.09.001

McFadden D (1974) The measurement of urban travel demand. J Public Econ 3(4):303–328. doi:10.1016/0047-2727(74)90003-6

Ophem HV, Stam P, Van Praag B (1999) Multichoice logit: modeling incomplete preference rankings of classical concerts. J Bus Econ Stat 17(1):117–128. doi:10.1080/07350015.1999.10524801

Rose JM (2014) Interpreting discrete choice models based on best-worst data: a matter of framing. In: Transportation research board 93rd annual meeting

Scarpa R, Notaro S, Louviere J, Raffaelli R (2011) Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. Am J Agric Econ 93(3):813–828

Tjur T (2009) Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. Am Stat 63(4):366–372. doi:10.1198/tast.2009.08210

Train KE (2009) Discrete choice methods with simulation. Cambridge University Press, Cambridge

Vermeulen B, Goos P, Vandebroek M (2011) Rank-order choice-based conjoint experiments: efficiency and design. J Stat Plan Inference 141(8):2519–2531. doi:10.1016/j.jspi.2011.01.019

Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. J Financ 16(1):8–37. doi:10.2307/2977633